Multi-Level Region Matching for Fine-Grained Sketch-Based Image Retrieval

Zhixin Ling
1069066484@qq.com
Fudan University
Shanghai, China

Zhen Xing
zxing20@fudan.edu.cn
Fudan University
Shanghai, China

Jiangtong Li
keep_moving-Lee@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Li Niu*
ustcnewly@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

ABSTRACT

Fine-Grained Sketch-Based Image Retrieval (FG-SBIR) is to use free-hand sketches as queries to perform instance-level retrieval in an image gallery. Existing works usually leverage only high-level information and perform matching in a single region. However, both low-level and high-level information are helpful to establish fine-grained correspondence. Besides, we argue that matching different regions between each sketch-image pair can further boost model robustness. Therefore, we propose Multi-Level Region Matching (MLRM) for FG-SBIR, which consists of two modules: a *Discriminative Region Extraction* module (DRE) and a *Region and Level Attention* module (RLA). In DRE, we propose Light-weighted Attention Map Augmentation (LAMA) to extract local feature from different regions. In RLA, we propose a transformer-based attentive matching module to learn attention weights to explore different importance from different image/sketch regions and feature levels. Furthermore, to ensure that the geometrical and semantic distinctiveness is well modeled, we also explore a novel LAMA overlapping penalty and a local region-negative triplet loss in our proposed MLRM method. Comprehensive experiments conducted on five datasets (*i.e.*, Sketchy, QMUL-ChairV2, QMUL-ShoeV2, QMUL-Chair, QMUL-Shoe) demonstrate effectiveness of our method.

CCS CONCEPTS

• **Computing methodologies** → *Matching*; Object recognition.

KEYWORDS

SBIR, FG-SBIR, Sketch-Based Image Retrieval, Sketch, Matching

ACM Reference Format:
Zhixin Ling, Zhen Xing, Jiangtong Li, and Li Niu. 2022. Multi-Level Region Matching for Fine-Grained Sketch-Based Image Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisbon, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548147

*Corresponding author

MM '22, October 10–14, 2022, Lisbon, Portugal.

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548147

1 INTRODUCTION

With the popularization of touch screen devices, abstract free-hand sketches have become an importance approach to interpret personalized needs of users. Thus, the research of Sketch-Based Image Retrieval (SBIR) has attracted increasing attention recently. The SBIR task can be categorized into Coarse-Grained category-level SBIR (CG-SBIR) [2, 6, 22] and Fine-Grained instance-level SBIR (FG-SBIR) [5, 19, 21, 25, 25, 29, 40, 48] by retrieval granularity. CG-SBIR retrieves an image in the gallery based on the category of query sketch while FG-SBIR retrieves a specific image that shares the same pose and outline. FG-SBIR has a wide range of applications in many fields such as searching online images or products.

Targeting at FG-SBIR, a number of existing works seeks to establish fine-grained correspondence by region matching: **1)** attention mechanisms[33, 48] explore to select a discriminative region to support region matching, but a single region might not able to learn good fine-grained correspondence; **2)** although part annotations are beneficial to part-level matching[16], such side information might be unavailable or expansive; **3)** recent works[26, 37] also divides images and sketches into grids to learn local correspondence. Whereas, a grid region might not contain semantically significant information. Besides, DLA[37] discovers that intermediate features are helpful to FG-SBIR. However, DLA simply discarded the last net block and did not take advantage of high-level information.

To overcome weaknesses of the existing works, we incorporate local information and low-level Convolutional Neural Networks(CNN) features into retrieval. Therefore, we propose multi-level region matching (MLRM) for FG-SBIR. MLRM first extracts features from different levels and regions by a *Discriminative Region Extraction* (DRE) module. Then a *Region and Level Attention* (RLA) module is adopted to learn attention weights to explore different contribution from different regions and levels.

In DRE, we propose a novel Light-weighted Attention Map Augmentation structure (LAMA) to extract discriminative attention maps. Inspired by CAMA[41], which adopts multiple network blocks for different regions, our LAMA reuses the same block to learn more generalizable discriminative regions. After that, we

explore a novel LAMA overlapping penalty and a local region-negative triplet loss to make different regions both geometrically and semantically discriminative. Furthermore, considering that low-level information (*e.g.*, texture, color and outlines) is helpful to establish fine-grained correspondence[20, 23], the attention maps obtained from LAMA are employed on both high-level and low-level feature maps to obtain region features.

In RLA, we convert region features into a sequence. Then a transformer encoder[34] is adopted to learn attention weights for different feature levels and image/sketch regions. Since the transformer is permutation-invariant, we construct a geometry map based on each region feature to learn different positional embeddings. At last, given a sketch-image pair, we compute feature distance of each level and region. The obtained distances are summed by learned attention weights as the retrieval distance. To distinguish some images that share similar parts with the query sketch, we also propose a novel global region-negative triplet loss.

Comprehensive experiments on five benchmark datasets [29, 42] (*i.e.*, Sketchy, QMUL-ChairV2, QMUL-ShoeV2, QMUL-Chair, QMUL-Shoe) have verified the effectiveness of our MLRM method. Our main contributions are summarized as follows: **1)** We propose a novel method named MLRM to perform multi-level region matching for FG-SBIR. **2)** We propose LAMA structure along with a LAMA overlapping penalty and a local region-negative triplet loss to extract geometrically and semantically discriminative regions. **3)** We design a novel transformer-based attentive matching approach, which is enhanced by novel geometry features. **4)** Our proposed MLRM achieves the best performance on four benchmark datasets.

2 RELATED WORK

Coarse-Grained SBIR. Coarse-Grained SBIR (CG-SBIR) was firstly proposed by Kato *et al.* [14], aiming to learn a feature space that closes the sketch-image domain gap. Existing CG-SBIR methods could be categorized into hand-crafted based methods and deep-learning based methods. Generally speaking, hand-crafted based methods first extracted edges maps from images and then designed hand-crafted features to match the query sketches with images [8, 10, 11, 28]. Deep-learning based methods using variants of siamese losses [33] and ranking losses [47] were proposed for the CG-SBIR task. Tu *et al.* [6] employ a two-stage training approach to better learn detailed features. A graph-based searching method was proposed in [2] to re-rank the retrieved images. Besides, CG-SBIR was also extended to the zero-shot setting [15, 31, 45] recently.

Fine-Grained SBIR. Fine-Grained SBIR (FG-SBIR) was firstly proposed by [18], which defined FG-SBIR as retrieving the image with the same attributes (*e.g.*, viewpoint and body configuration) as the query sketch. Yu *et al.* [42] extended the definition, requiring the retrieved image to be right corresponding to the query sketch. Since attribute annotations might be unavailable, we follow the definition in [42] like most of the existing works [5, 7, 19, 25, 27, 39]. Apart from the huge sketch-image domain gap, FG-SBIR also needs to capture fine-grained correspondence between images and sketches, which makes FG-SBIR more challenging.

Qian *et. al.* introduced sketch augmentation approaches by stroke deformation and stroke removal. Various classification losses were proposed in [19, 29] to close the sketch-image domain gap. Seddati *et al.* [30] explored a quadruplet loss to regulate both intra-class and inter-class distances. To perform shape matching, images were converted into edge maps to overcome the domain gap in [27]. To enhance cross-domain generalization ability, several works introduced text-domain triplet losses [12, 40]. Bhunia *et. al.* [4] sought to detect noisy stroke via a reinforcement learning approach.

When the above methods manipulated global features of images and sketches for matching, several works investigated local features for FG-SBIR. By taking advantage of part labels, Li *et. al.* [18] utilized a deformable part-based model [9] and a graph matching model to perform part-level matching. A reinforcement learning method was proposed in [5], in which the sketch representation at each rendering step is rewarded by retrieving the paired photo early. Li *et al.* [16] and Li *et al.* [17] proposed part-aware models for part detection, which located object parts with a strongly-supervised deformable part-based model [1]. Although these works[5, 9, 16–18] worked well in learning local features, the required side information (*e.g.*, part labels, stroke annotations or instance attributes) can be unavailable or expansive in practice.

Attention mechanisms [33, 48] were introduced to learn fine-grained correspondence, but only focused on a single region. Some recent works[26, 37] divided images/sketches into grids to learn local correspondence. Pang *et. al.* designed a multi-modal jigsaw puzzle [26]. Xu *et. al.*[37] proposed DLA that matches each sketch pixel with an image pixel on the feature map. But simple grids might not be semantically significant, probably leading to performance bottleneck.

Compared with previous works, MLRM extracts multiple semantically discriminative regions without side information. Moreover, we leverage both high-level and low-level features to learn fine-grained correspondence.

Deep Discriminative Region Discovery. Without utilizing part-level annotations, discriminative regions can be discovered by attention maps[36, 38, 48] or class activation maps[24, 41, 46]. Most of existing works focused on a single discriminative region [38, 48] while others [35, 44] could discover multiple regions. Approaches based on adversarial erasing [24, 35, 44] learned multiple discriminative regions in series. In contrast, Class Activation Maps Augmentation (CAMA) [41] learned multiple activation maps in parallel. However, CAMA used multiple branches and consumed a large number of model parameters. The required class labels in CAMA were also sometimes unavailable. Nevertheless, our LAMA is light-weighted and able to learn multiple regions in parallel.

3 OUR METHOD

Our proposed Multi-Level Region Matching model (MLRM) consists of two modules: a *Discriminative Region Extraction* module (DRE) and a *Region and Level Attention* module (RLA). The overview of MLRM is shown in Fig. 1. DRE extracts discriminative regions based multi-level CNN features by a novel LAMA structure (Fig. 2). Then RLA utilizes a transformer encoder to convert multi-level region features into region weights and level weights. In RLA, we also propose to construct a geometry map to obtain positional embeddings. At last, given a sketch-image pair, we can derive retrieval distance using their region features and the obtained weights. The workflow is the same for both sketches and images if not specified. When

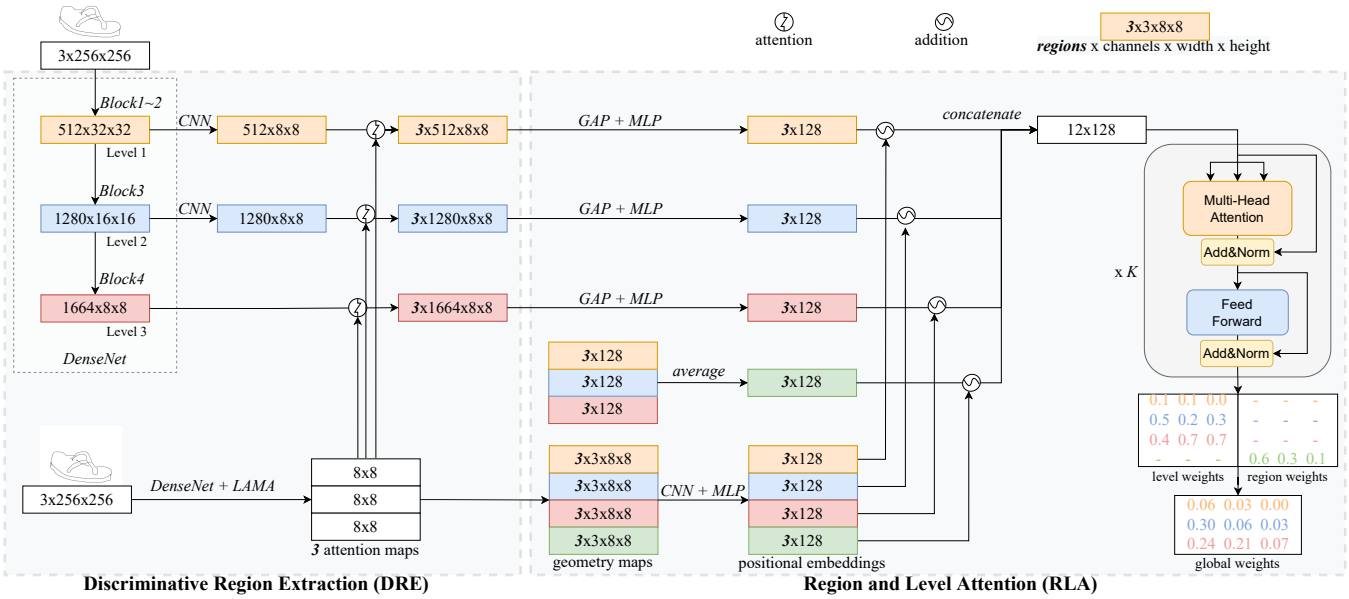

Figure 1: An overview of MLRM. The two *DenseNet* in DRE are the same one.

introducing the retrieval distance (Sec. 3.2.3), we will distinguish sketches and images by introducing two symbols s/p.

3.1 Discriminative Region Extraction

3.1.1 Multi-Level Feature Map Extraction. Following TC-Net[19], DRE adopts DenseNet169[13] as backbone. DenseNet169 mainly consists of 4 dense blocks and 3 transition blocks. The input sketch or image is resized to 256×256[37]. We take advantage of intermediate outputs of the last 3 dense blocks, denoted by $\mathbf{F}'_1 \in \mathbb{R}^{512\times32\times32}, \mathbf{F}'_2 \in \mathbb{R}^{1280\times16\times16}, \mathbf{F}'_3 \in \mathbb{R}^{1664\times8\times8}$. Intuitively, $\mathbf{F}'_1$ contains abundant low-level information (*e.g.*, color, shape, and outlines) while $\mathbf{F}'_3$ contains abundant high-level information (*e.g.*, poses, categories and attributes) [20, 23]. DLA[37] shows that $\mathbf{F'}_2$ is the best for establishment of fine-grained correspondence and simply removes the last CNN block. This is because $\mathbf{F}_2$ strikes a balance between low-level and high-level information. However, our method consider all these three levels. In order to unify the feature map sizes, we adopt two CNNs to reduce $\mathbf{F}'_1$ into $\mathbf{F}_1 \in \mathbb{R}^{512\times8\times8}$ and $\mathbf{F}'_2$ into $\mathbf{F}_2 \in \mathbb{R}^{1280\times8\times8}$. For naming consistency, we also denote $\mathbf{F}'_3$ as $\mathbf{F}_3$. We call $\mathbf{F}_l$ as Level$l$ feature map, $l \le L, L = 3$.

3.1.2 The LAMA Approach to Extract Different Attention Maps. Inspired by Class Activation Map Augmentation (CAMA)[41] that adopts three branches of network blocks to produce class activation maps, we propose Light-weighted Attention Map Augmentation (LAMA) to derive spacial attention maps to attend the obtained multi-level feature maps, as shown in Fig. 2. LAMA learns an indicator map from a one-hot indicator vector $v_r \in \mathbb{R}^N$ to supplement intermediate feature maps of the backbone. $v_r \in \mathbb{R}^N$ is a pre-defined input. We set the r-th element to 1 to obtain the r-th attention map. After that, the backbone output along with $\mathbf{F}_2$ and $\mathbf{F}_1$ is used to generate an attention map $\mathbf{M}_r \in \mathbb{R}^{1\times8\times8}$ through an attention CNN E^{LAMA}. The involvement of $\mathbf{F}_2$ and $\mathbf{F}_1$ is necessary because the attention map aims to attend multi-level feature maps. In particular, the procedure can be explained in formla:

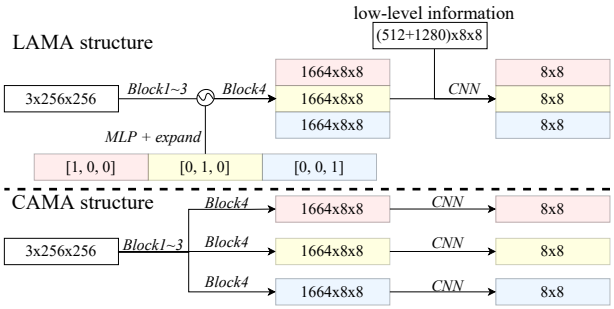

Figure 2: LAMA and CAMA structure.

$$
\begin{aligned}
\mathbf{F}_3^r &= Block4(E^v(v_r) + \mathbf{F}'_2); \\
\mathbf{M}'_r &= E^{LAMA}([AvgPool(\mathbf{F}_3^r); MaxPool(\mathbf{F}_3^r); AvgPool(\mathbf{F}_2); \\
&\qquad MaxPool(\mathbf{F}_2); AvgPool(\mathbf{F}_1); MaxPool(\mathbf{F}_1)]); \\
\mathbf{M}_r &= sigmoid(\mathbf{M}_r - m_r^\tau),
\end{aligned} \tag{1}
$$

where $r = 1, 2, \cdots, N$ and N is the number of attention maps; E^v contains fully connected (fc) layers with ReLU activation, transforming v_r into a vector of $\mathbb{R}^{1280}$ and then expanding it into $\mathbb{R}^{1280\times16\times16}$; $Block4$ is the 4-th dense block of DenseNet169; $AvgPool/MaxPool$ is channel-wise average/max pooling[36]; [...] is channel-wise concatenation; m_r^τ is the τ-th largest element of $\mathbf{M'}_r$. In vector v_r, only the r-th element is 1 and the others are 0. One attention map locates one local region. τ decides the region size. We empirically set $N = 3$ and $\tau = 3$. It is a good way to generate multiple good-quality attention map by taking advantage of $Block4$, which is able to leverage semantic information learned by $Block4$.

Compared with CAMA that makes N copies of *Block*4, LAMA saves a large number of model parameters simply by using an indicator vector. Moreover, LAMA requires no class labels and can be accustomed to FG-SBIR. After that, we can obtain the *multi-level region features* (*region features* in short): $\mathbf{f}_{r,l} = GAP(\mathbf{M}_r \odot \mathbf{F}_l)$, where $r \le N, l \le L$ and GAP is Global Average Pooling [20].

3.2 Region and Level Attention

3.2.1 Derivation of Gometry Maps and Positional Embeddings. Different regions and levels may contribute differently to the matching procedure, so we introduce a *Region and Level Attention* module (RLA) to attend the important regions and levels. Considering that prediction of these attention weights should capture relations among all regions and levels. We therefore adopt a transformer encoder for this purpose. We assumes that region weights $\mathbf{W}_N$ are level-agnostic while level weights $\mathbf{W}_L$ are region-dependent. So the goal of RLA is to learn two matrix: $\mathbf{W}_N \in \mathbb{R}^{1\times N}, \mathbf{W}_L \in \mathbb{R}^{L\times N}$. First we utilize L different fc layers to reduce $\mathbf{f}_{r,l}$ into the same dimension d. Let the reduced features be denoted as $\mathbf{f}'_{r,l}$. To obtain $\mathbf{W}_N$, we derive a feature averaging reduced features of all levels. For simplicity, let the averaged feature belongs to Level $L+1$: $\mathbf{f}'_{r,L+1}(i) = \frac{1}{L}\sum_l \mathbf{f}'_{r,l}(i)$. Since the transformer architecture is permutation-invariant, a positional embedding is required for each reduced feature. Considering that the positional embeddings should be aware of both the level and region, we construct a geometry map $\mathbf{G}_{r,l} \in \mathbb{R}^{L\times 8\times 8}$ from attention maps:

$$\mathbf{G}_{r,l}(i,j,k) = \begin{cases} \mathbf{M}_r(j,k), & l \in \{L+1, i\}; \\ 0, & otherwise. \end{cases} \tag{2}$$

When $1 \le l \le L$, we only allow channel l to specify the attended region. Since features of level $L+1$ is an average of the previous levels, we let all channels specify the attended region. Then a geometry encoder E^G consisting of CNN and MLP is adopted to encode $\mathbf{G}_{r,l}(i,j,k)$ into d-dimensional positional embeddings. We add positional embeddings to $\mathbf{f}'_{r,l}$ to retain positional information. Let the obtained embeddings be denoted as $\mathbf{f}''_{r,l} \in \mathbb{R}^d$.

3.2.2 Attention Weight Generation. The transformer encoder expects a sequence as input, so we concatenate $\mathbf{f}''_{r,l}$ into a sequence $\mathbf{S} \in \mathbb{R}^{(L+1)N\times d}$, where $\mathbf{S}(Nl - N + r, \cdot) = \mathbf{f}''_{r,l}$. Two fc heads are inserted after the transformer encoder to obtain *Region and Level Attention* weights. Let E^t stand for the transformer encoder and these fc heads, we can formulate the procedure: $(\mathbf{W}_N, \mathbf{W}_L) = E^t(\mathbf{S})$. $\mathbf{W}_N$ corresponds to the last N outputs, *i.e.*, $\mathbf{S}(i \ge NL - N + 1, \cdot)$. $\mathbf{W}_L$ corresponds to the first $NL - N$ outputs, *i.e.*, $\mathbf{S}(i \le NL - N, \cdot)$. At last, we can derive a global weight matrix $\mathbf{W} \in \mathbb{R}^{L\times N}$:

$$\begin{aligned} \mathbf{W}'_L(\cdot, r) &= softmax(\mathbf{W}_L(\cdot, r)), \quad r \le N; \\ \mathbf{W}(l, \cdot) &= softmax(\mathbf{W}_N(1, \cdot)) \odot \mathbf{W}'_L(l, \cdot), \quad l \le L. \end{aligned} \tag{3}$$

3.2.3 Retrieval via Attentive Matching. Having acquiring attention wights for each region and level, we compute a global retrieval distance between a given sketch s and a given image p. Taking s/p as superscript for related mathametical representation, *i.e.*, region features $\mathbf{f}^s_{r,l}/\mathbf{f}^p_{r,l}$ and global weight matrices $\mathbf{W}^s/\mathbf{W}^p$. If the p and s are unpaired or not well aligned, their region-level weights can be different. We sum the distances between $\mathbf{f}^s_{r,l}$ and $\mathbf{f}^p_{r,l}$ by an averaged attention weights as the retrieval distance D':

$$D'(s,p) = \frac{1}{2}\sum_{r,l}(\mathbf{W}^s(l,r) + \mathbf{W}^p(l,r)) \times D(\mathbf{f}^s_{r,l}, \mathbf{f}^p_{r,l}). \tag{4}$$

3.3 Loss Functions

3.3.1 Overlapping Penalty. Overlapped regions are insignificant for region matching. A CAMA solution is to employ an overlapping penalty on the attention maps [41], which is formulated as:

$$\mathcal{L}_{ovl-cama} = \frac{1}{N}\sum_{x,y} \mathbf{M}_1 \odot \mathbf{M}_2 \odot \cdots \odot \mathbf{M}_N, \tag{5}$$

where x, y are coordinates along X-axis and Y-axis respectively and $\odot$ is element-wise multiplication. Intuitively, $\mathcal{L}_{ovl-cama}$ penalize the regions attended by all attention maps and enforce different attention maps to locate different regions. However, since inactivated pixels are close to zero, the penalty will be weakened in the regions where the pixels are inactivated in some attention map. Therefore, we propose a LAMA overlapping penalty:

$$\mathcal{L}_{ovl-lama} = \frac{1}{N\times N}\sum_{r\le N}\sum_{x,y} \mathbf{M}_r \odot MaxPool(\prod_{r'\ne r, r'\le N} \mathbf{M}_{r'}), \tag{6}$$

where $\prod$ is continual channel-wise concatenation. For each attention map $\mathbf{M}_r$, $\mathcal{L}_{ovl-lama}$ first aggregates activated regions of other attention maps by $MaxPool$, and then reduces the overlapping between $\mathbf{M}_r$ and the aggregated attention map. The overlapping penalty is able to make the regions geometrically discriminative.

3.3.2 Local Triplet Loss. We adopt a local triplet loss $\mathcal{L}_{ltrp}$ to make region features in each level semantically discriminative:

$$\begin{aligned} \mathcal{L}_{ltrp-in} &= max(D(\mathbf{f}_{r,l}, \mathbf{f}^+_{r,l}) - D(\mathbf{f}_{r,l}, \mathbf{f}^-_{\cdot,l}) + (1 + e^{-l})m, 0); \\ \mathcal{L}_{ltrp-rn} &= max(D(\mathbf{f}_{r,l}, \mathbf{f}^+_{r,l}) - D(\mathbf{f}_{r,l}, \mathbf{f}^+_{r_-,l}) + (1 + e^{-l})\alpha m, 0); \\ \mathcal{L}_{ltrp} &= \mathcal{L}_{trp-in} + \mathcal{L}_{trp-rn}, \end{aligned} \tag{7}$$

where $\mathbf{f}_{r,l}$ is the anchor feature; $\mathbf{f}^+_{r,l}$ is the positive feature; $\mathbf{f}^-_{\cdot,l}$ an instance-negative feature and $\cdot$ can be any region; $\mathbf{f}^+_{r',l}$ is a region-negative feature; m is the triplet margin and α is a scale parameter ranging from 0 to 1; D is a distance metric, usually L2 distance. In FG-SBIR, the anchor is a sketch while positive/negative samples are images. There are two terms in $\mathcal{L}_{ltrp}$. We refer to $\mathcal{L}_{ltrp-in}/\mathcal{L}_{ltrp-rn}$ as *local instance-negative/region-negative triplet loss*: **1)** The first term $\mathcal{L}_{ltrp-in}$ is similar to common triplet losses[19, 26, 29, 37] that pull the positive feature closer to the anchor when pushing negative feature away. Because low-level information (*e.g.* outlines) should be more diversified than high-level information (*e.g.* categories), we relax the margin by a coefficient $1 + e^{-l}$. **2)** Besides, although $\mathcal{L}_{ovl-lama}$ make regions geometrically different, it might not ensure that different regions are semantically different and that the same regions are semantically similar. In this case, the region matching might not work well. So we propose the loss $\mathcal{L}_{ltrp-rn}$ that trys to push region-negative feature $\mathbf{f}^+_{r_-,l}$ away by a tighter margin αm. With both $\mathcal{L}_{ovl-lama}$ and $\mathcal{L}_{ltrp-rn}$, we can extract both geometrically and semantically discriminative region features.

	Sketchy (%)	QMUL-ChairV2 (%)	QMUL-ShoeV2 (%)	QMUL-Chair (%)	QMUL-Shoe (%)
Song *et al.* [32] (CVPR '16)	-	-	-	78.4	50.4
GN Triplet [29] (TOG '16)	37.1	-	-	-	-
SaN Triplet [42] (CVPR '16)	36.2	56.6	30.9	72.2	52.2
Quadruplet [30] (ACM MM '17)	42.2	-	-	-	-
DSSA [33] (ICCV '17)	-	-	33.7	81.4	61.7
Radenovic *et al.* [27] (ECCV '18)	-	-	-	85.6	54.8
DCCRM [40] (PR '19)	46.2	-	-	-	-
TC-Net [19] (ACM MM '19)	40.8	65.3	40.2	95.9	63.5
Bhunia *et al.* [5] (CVPR '20)	-	(89.7)	(79.6)	-	-
Pang *et al.* [26] (CVPR '20)	-	-	36.5	96.0	56.5
Bhunia *et al.* [3] (CVPR '21)	-	60.2	39.1	-	-
LA [37] (ACM MM '21)	43.1	64.8	42.3	93.8	57.4
DLA [37] (ACM MM '21)	54.9	69.2	50.2	**99.0**	**79.1**
Zhang *et al.* [43] (PR '22)	-	-	-	84.4	65.7
AE-Net [7] (PR '22)	46.0	-	-	-	-
Bhunia *et al.* [4](CVPR '22)	-	64.8	43.7	-	-
MLRM (ours)	**57.2**	**74.3(98.2)**	**50.4(87.9)**	**99.0**	67.0

Table 1: acc@1(acc@10) comparison with previous works.

		TC-Net[19]	LA [37]	DLA [37]	MLRM (ours)
QMUL-ChairV2	Time (s)	5.3	27.5	236.7	11.8
	acc@1 (%)	65.3	64.8	69.2	74.3
Sketchy	Time (s)	8.2	46.8	639.3	14.1
	acc@1 (%)	40.8	43.1	54.9	57.2

Table 2: Retrieval time comparison using the same GPU.

3.3.3 Global Triplet Loss. To optimize $\mathbf{W}^s$ and $\mathbf{W}^p$, a global triplet loss is employed on $D'(s,p)$:

$$\begin{aligned}\mathcal{L}_{gtrp-in} &= max(D'(s,p^+) - D'(s,p^-) + m, 0);\\ \mathcal{L}_{gtrp-rn} &= max(D'(s,p^+) - D'(s,p^+_{r'}) + \beta m, 0);\\ \mathcal{L}_{gtrp} &= \mathcal{L}_{gtrp-in} + \mathcal{L}_{gtrp-rn},\end{aligned} \tag{8}$$

where p^+/p^- is paired/unpaired with s; β ranges from 0 to 1, similar to α in Eq. 7; $p^+_{r'}$ stands for a constructed region-negative image, $r' \le N$. In each level, we replace the r'-th region feature of p^+ with that of p^- to construct $p^+_{r'}$. Specifically, let $\mathcal{S}_p = \{\mathbf{f}^p_{r,\cdot} | r \le N\}$ be the feature set generated from p. Then $p^+_{r'}$ corresponds to feature set $\mathcal{S}_{p^+_{r'}} = \{\mathbf{f}^{p^+}_{r,\cdot} | r \ne r', r \le N\} \cup \{\mathbf{f}^{p^-}_{r,\cdot} | r = r'\}$. $\mathbf{W}^{p^+_{r'}}$ is derived from $\mathcal{S}_{p^+_{r'}}$. $D'(s,p^+_{r'})$ is calculated based on $\mathcal{S}_{p^+_{r'}}$ and $\mathcal{S}_s$ in Eq. 4. We only replace the region feature, and do not replace the positional embeddings. Then there is only one unpaired region between $p^+_{r'}$ and s, so we employ a tighter triplet margin βm in $\mathcal{L}_{gtrp-rn}$. Unlike the common $\mathcal{L}_{gtrp-in}$, $\mathcal{L}_{gtrp-rn}$ enables our model to distinguish the negative images that might share some similar parts with s. We refer to $\mathcal{L}_{gtrp-in}/\mathcal{L}_{gtrp-rn}$ as *global instance-negative/region-negative triplet loss.*

3.3.4 The Total Loss. At last, we can derive the total loss:

$$\mathcal{L}_{total} = \omega_{gtrp}\mathcal{L}_{gtrp} + \omega_{ltrp}\mathcal{L}_{ltrp} + \mathcal{L}_{ovl-lama}, \tag{9}$$

where ω_{gtrp} and ω_{ltrp} are hyper-parameters that balance the weight among different loss terms.

4 EXPERIMENTS

4.1 Experiment Setup

4.1.1 Dataset. We evaluate our MLRM method on five benchmark fine-grained sketch-image datasets:

Sketchy [29] is the largest fine-grained sketch-image dataset, totally containing 74425 sketches and 12500 images from 125 categories. Each image is paired with at least five sketches. Images in the Sketchy dataset can be quite noisy and the paired sketches can also be misaligned (*e.g.*, the sketch can be scaled, rotated, or distorted). Besides, there could be several very similar images. These factors makes this dataset challenging. Following [29], 90% images and their paired sketches are used for training and the rest for testing. To take advantage of the class labels in Sketchy, instead of inserting a classification head[37], we fill in a batch with sketch-image pairs from different categories and the same category alternately.

QMUL-Shoe and **QMUL-Chair** [42] contain 419 and 297 sketch-image pairs respectively, where each image is paired with only one sketch. All the images and sketches in these two datasets are clean and well aligned. Following the split of [42], 304/200 pairs of QMUL-Shoe/QMUL-Chair are used for training and the rest for testing.

QMUL-ShoeV2 and **QMUL-ChairV2** [42] are extensions of QMUL-Shoe and QMUL-Chair respectively. Each image is paired with at least three sketches in these two datasets. Similarly, images and sketches in this dataset are clean and well aligned. There are totally 2000/400 images and 6730/1275 sketches in the QMUL-ShoeV2/QMUL-ChairV2. Following [19], we use 1800/300 images and their paired sketches for training and the rest for testing.

4.1.2 Implementation Details. Please refer to Supplementary at https://github.com/1069066484/MLRM-ACMMM2022.

4.2 Comparison with Previous Works

4.2.1 Accuracy comparison. We compare MLRM with 16 previous baselines and report the results in Tab. 1. On large datasets (*i.e.*, Sketchy, QMUL-ChairV2, QMUL-ShoeV2), MLRM consistently

setting	alternatives of MLRM	Sketchy	QMUL-ChairV2	QMUL-ShoeV2	QMUL-Chair	QMUL-Shoe
S1	single region ($N = 1$)	47.3	67.9	41.8	95.9	49.6
S2	$w/o\ \mathcal{L}_{ovl-lama}$	46.1	66.2	40.7	92.8	47.8
S3	replace $\mathcal{L}_{ovl-lama}$ with $\mathcal{L}_{ovl-cama}$	55.0	68.8	47.0	96.9	59.1
S4	single level (only enable Level3)	54.8	57.6	44.1	85.6	43.5
S5	w/o geometry maps ($f''_{r,l} \leftarrow f'_{r,l}$)	55.0	72.4	48.7	86.0	61.7
S6	$w/o\ \mathcal{L}_{gtrp}$ ($D'(s,p) \leftarrow \frac{1}{N\times L}\sum_{r,l} D(\mathbf{f}^s_{r,l}, \mathbf{f}^p_{r,l})$)	42.5	55.3	36.8	88.7	47.0
S7	$w/o\ \mathcal{L}_{ltrp}$	52.5	70.0	43.8	94.8	57.4
S8	$w/o\ \mathcal{L}_{ltrp-rn}$ ($\mathcal{L}_{ltrp} \leftarrow \mathcal{L}_{ltrp-in}$)	53.9	71.1	45.1	95.9	59.1
S9	$w/o\ \mathcal{L}_{gtrp-rn}$ ($\mathcal{L}_{gtrp} \leftarrow \mathcal{L}_{gtrp-in}$)	55.9	72.9	48.8	97.9	62.6
S10	**full model**	**57.2**	**74.3**	**50.4**	**99.0**	**67.0**

Table 3: Ablation results.

outperforms previous works, among which DLA[37] is the best baseline. We have the following observations: **1)** On Sketchy, the largest and most diversified dataset, MLRM outperforms DLA by 2.3% and the second best by 11.0%. On QMUL-ChairV2/QMUL-ShoeV2, two large clean and aligned datasets, MLRM outperforms DLA by 5.1%/0.2% and the second best by 9.0%/6.7%. **2)** On two small datasets (*i.e.*, QMUL-Chair, QMUL-Shoe), MLRM outperforms previous works beside DLA. On QMUL-Chair, MLRM has equivalent acc@1 as DLA and outperforms the second best baseline by 3.1%. On QMUL-Shoe, MLRM result is 12.1% lower than DLA but outperforms the second best baseline by 1.3%. These results validates that MLRM is effective and favors large datasets.

4.2.2 Retrieval Time Comparison. Tab 1 shows that DLA[37] is the strongest baseline. Thus, we further compare retrieval time of MLRM, DLA and several other representative works in Tab. 2. In general, the ranking of retrieval time is: DLA $\gg$ LA $\gg$ MLRM $>$ TCNet. Although DLA achieves impressive performances, the retrieval time is much longer than other works. TCNet retrieval pipeline [19] is typical for most of works[29, 30, 42]. Given features from a sketch and an image, TCNet simply computes a distance. MLRM need compute $N \times L$ distances from pairs of region features. The computation of MLRM is therefore about 9 times of TCNet, where $9 = N \times L$. However, compared with the performance gain, the extra computation overhead is acceptable.

In contrast, LA[37] computes the distance on the feature map pixel by pixel. DLA[37] slides each sketch pixel over the whole image to locate the matched image pixel. Let the feature map size be 16×16, the computation overhead of LA/DLA is about $16^2/256^2$ times of TCNet. Despite the competitive performance of DLA, MLRM certainly has good advantage in retrieval time (*e.g.*, DLA 639.3s *v.s.* MLRM 14.1s on Sketchy).

4.3 Ablation Study

We do ablation study for MLRM and report the results in Tab. 3. Setting S1-3 show the important role of region matching in MLRM. Matching through a single region (S1) leads to poor performance (S1 *v.s* S10). Moreover, overlapping penalty is crucial for region matching. Without it, both of $\mathcal{L}_{ltrp-rn}$ and $\mathcal{L}_{gtrp-rn}$ would make no sense, therefore leading to sharp performance drop (S2 *v.s* S10). S3 validates that $\mathcal{L}_{ovl-lama}$ prevents different regions from overlapping more efficiently than $\mathcal{L}_{ovl-cama}$ does.

On Sketchy, misalignment of paired sketches and images results in more mistakes in low-level matching. Different categories also improve importance of Level3 information. Therefore, only enabling Level3 (S4) make the Sketchy result drop by 2.4% while results on other datasets drop much more sharply (at least 6.3%). S5 shows that our positional embeddings can effectively supplement geometry information for the RLA transformer encoder E^t. The information is helpful to learn $\mathbf{W}_N$ and $\mathbf{W}_L$ and then improve model performances. S6 implies that different regions and different levels should have different importance. A simple average of region feature distance $D(\mathbf{f}^s_{r,l}, \mathbf{f}^p_{r,l})$ leads to a sharp performance drop.

$\mathcal{L}_{ltrp}$ can close the distance between paired region features, *i.e.* $\mathbf{f}^s_{r,l}$ and $\mathbf{f}^p_{r,l}$. Ablating it is harmful to model performance (S7). S8 reveals that $\mathcal{L}_{ltrp-rn}$ can further improve region feature quality by enforcing different regions to be semantically discriminative. Through $\mathcal{L}_{gtrp-rn}$, more negative images that resembles the sketch are fed into the RLA transformer encoder E^t. It improves model performance by at least 1.0% acc@1 on various datasets.

4.4 Further Study

4.4.1 Better Discriminative Regions: LAMA v.s. *CAMA.* DRE produces discriminative regions via LAMA. We study LAMA by comparing it with CAMA[41]. We first visualize attention maps in Fig. 3 to show effectiveness of two related losses, *i.e.*, $\mathcal{L}_{ovl-lama}$ and $\mathcal{L}_{ltrp-rn}$. We have the following observations: **1)** Compared with $\mathcal{L}_{ovl-cama}$, our proposed $\mathcal{L}_{ovl-lama}$ prevents different regions from overlapping more effectively. With $\mathcal{L}_{ovl-cama}$ (Fig. 3-(1)), it turns out that two attention maps focus on the same region, *i.e.*, the chair bottom. In contrast, $\mathcal{L}_{ovl-lama}$ makes the three attention maps focus on the top, middle and bottom of the chair respectively (Fig. 3-(3)). In other words, $\mathcal{L}_{ovl-lama}$ can effectively learn geometrically discriminative attention maps. **2)** Our proposed local region-negative triplet loss $\mathcal{L}_{ltrp-rn}$ is able to enforce the extracted regions to be semantically discriminative. The three attention maps focus on different parts of the chair (Fig. 3-(3)). On the contrary, without $\mathcal{L}_{ltrp-rn}$, two attention maps locate feet of the chair ($\mathbf{M}_1$ and $\mathbf{M}_2$ in Fig. 3-(4)), which might not help much to region matching. **3)** Without overlapping penalty, all attention maps tend to focus on

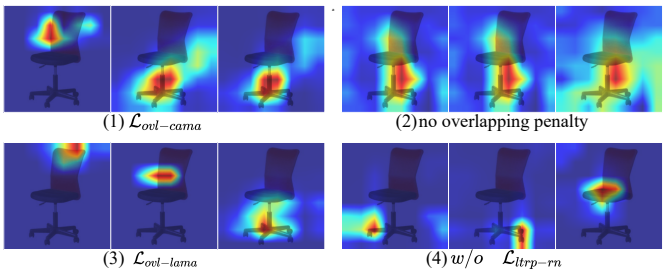

Figure 3: Visualization of attention maps $\mathbf{M}_r$.

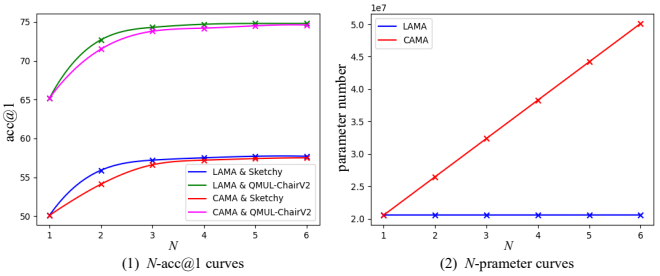

Figure 4: LAMA and CAMA quantitative comparison.

strategy	Definition of $D'(s,p)$	Sketchy	QMUL-ChairV2
T1	$\sum_{r,l} \mathbf{W}^s(l,r) \times D(\mathbf{f}^s_{r,l}, \mathbf{f}^p_{r,l})$	56.4	73.2
T2	$\sum_{r,l} \mathbf{W}^p(l,r) \times D(\mathbf{f}^s_{r,l}, \mathbf{f}^p_{r,l})$	56.1	73.0
T3	$\mathbb{E}_r D(\mathbf{f}^s_{r,1}, \mathbf{f}^p_{r,1})$	28.1	51.1
T4	$\mathbb{E}_r D(\mathbf{f}^s_{r,2}, \mathbf{f}^p_{r,2})$	50.7	70.7
T5	$\mathbb{E}_r D(\mathbf{f}^s_{r,3}, \mathbf{f}^p_{r,3})$	54.0	69.4
T6	$\mathbb{E}_l \sqrt{\sum_r min_{r'} D(\mathbf{f}^s_{r,l}, \mathbf{f}^p_{r',l})^2}$	45.2	64.9
T7	$\mathbb{E}_{r,l} D(\mathbf{f}^s_{r,l}, \mathbf{f}^p_{r,l})$	46.9	66.5
T8	$\mathbb{E}_{r,l} min_{r'} D(\mathbf{f}^s_{r,l}, \mathbf{f}^p_{r',l})$	47.0	66.5
T9	Eq. 4 (ours)	57.2	74.3

Table 4: Retrieval strategy comparison.

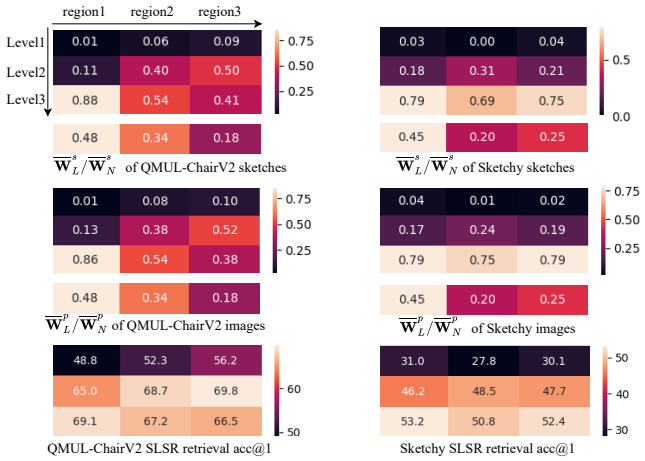

Figure 5: Visualization of average attention weights (*e.g.*, $\overline{\mathbf{W}}^s_L$ and $\overline{\mathbf{W}}^s_L$) and SLSR performance.

the same region (Fig. 3-(2)). Although $\mathcal{L}_{ltrp-rn}$ tries to make the regions semantically discriminative, it does not directly manipulate the attention maps. Therefore, it fails to learn the attention maps alone without $\mathcal{L}_{ovl-lama}$.

The difference between LAMA and CAMA network structures is shown in Fig 2. In addition, we show their quantitative comparison in Fig. 4. For fair comparison, we also supplement low-level information in CAMA like the derivation of $\mathbf{M}'_r$ in Eq. 1. On the one hand, Fig. 4-(1) shows that **1)** LAMA consistently outperforms CAMA, which is probably because reusing of *Block*4 in LAMA can effectively boost generalization ability; **2)** LAMA performance gain becomes insignificant when $N \geq 4$, implying that setting $N = 3$ is enough for region matching. On the other hand, since the last net block (*e.g. Block*4 of DenseNet) is the most parameter-consuming, LAMA saves a large number of parameters by merging N branches into one. In contrast, the parameter number of CAMA linearly increases quickly as N grows (Fig. 4-(2)).

4.4.2 Effectiveness of Attentive Matching. To further study RLA, in Fig. 5, we visualize mean of the attention weights (*e.g.*, $\overline{\mathbf{W}}^s_L(l,r) = \mathbb{E}_s \mathbf{W}^s_L(l,r)$) and single-level single-region (SLSR) retrieval performance. There is a difference between results on these two datasets: compared with Sketchy, QMUL-ChairV2 relatively prefers Level2. There might be two reasons: **1)** Sketchy contains objects from various categories and needs high-level features for retrieval; **2)** sketches and images of QMUL-ChairV2 are strictly aligned, so the matching can focus more on low-level features and care less about high-level features. Our conjecture can also be validated by SLSR performances: Level3 retrieval on Sketchy obviously outperforms Level2 while retrieval results in these two levels are close on QMUL-ChairV2. SLSR comparison also shows that different feature levels and image/sketch regions result in different retrieval performances. Thus it is important to learn good attention weights.

There are also three common observations on both datasets: **1)** For $\mathbf{W}^{\cdot}_L$: $\mathbf{W}^{\cdot}_L(1,\cdot) < \mathbf{W}^{\cdot}_L(2,\cdot) < \mathbf{W}^{\cdot}_L(3,\cdot)$, implying that high-level features always helps the matching procedure; **2)** the region weights $\mathbf{W}^{\cdot}_N$ are almost the same for both sketches and images, probably implying that the extracted regions are semantically significant and do not simply capture low-level information (*e.g.* outlines and texture) that varies from one image/sketch to another; **3)** There is not much difference between $\mathbf{W}^s_L$ and $\mathbf{W}^p_L$. Therefore we further compute the standard deviation of each weight and find out that $Std(\mathbf{W}^{\cdot}_L(l,r)) \leq 0.06$ for each $l \leq L, r \leq N$, showing that the attention weights for different images and sketches do not vary very much. However, it does not indicate that combination of sketch weights and image weights (Eq. 4) is helpless. We compare different retrieval strategies in Tab. 4. T1/T2 shows that using only sketch/image weights leads to 0.8%/1.1% performance drop on Sketchy and 1.1%/1.3% performance drop on QMUL-ChairV2. So weight combination of Eq. 4 is helpful.

Tab. 4 also reveals that retrieval in a single level (T3-T5) does not leads to good performance. Moreover, it shows importance of different levels depends on different datasets: **1)** on Sketchy, Level3

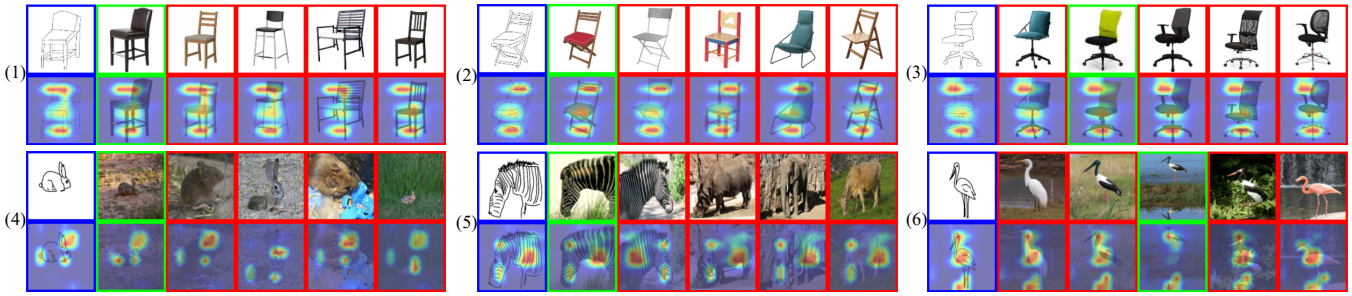

Figure 6: Top-5 retrieval visualization on QMUL-ChairV2(row (1)-(3)) and Sketchy(row (4)-(6)). The sketches bordered in blue are queries. The images bordered in green/red are positive/negative cases.

		Sketchy	QMUL-ChairV2
Level1	$w/o\ \mathcal{L}_{ltrp-rn}$	0.157	0.116
	$w\ \mathcal{L}_{ltrp-rn}$	3.675	6.540
Level2	$w/o\ \mathcal{L}_{ltrp-rn}$	0.136	0.107
	$w\ \mathcal{L}_{ltrp-rn}$	2.574	6.297
Level3	$w/o\ \mathcal{L}_{ltrp-rn}$	0.257	0.162
	$w\ \mathcal{L}_{ltrp-rn}$	2.882	3.924

Table 5: $\overline{D}_l$ comparison.

> Level2 > Level1; **2)** on QMUL-ChairV2, Level2 $\gtrsim$ Level3 > Level1. Because Level1 retrieval results in the worst performance on both datasets, we train MLRM by ablating Level1 to study its necessity. The resultant acc@1 drops by 0.4%/0.9% on Sketchy/QMUL-ChairV2. Therefore, Level1 is beneficial for MLRM.

Besides, we conduct retrieval using DLA-style region distance (T6). DLA[37] slides each sketch pixel over the image feature map, while T6 is like sliding each sketch region over all image regions. Results show that T6 is inferior to our attentive matching (T9). Besides, we also find that $T6$ is close to $T7$. We conjecture that it is because there exists a equivalent relation between $\mathbf{f}^s_{r,l}$ and $\mathbf{f}^p_{r,l}$: $min_{r'} D(\mathbf{f}^s_{r,l}, \mathbf{f}^p_{r',l}) \approx D(\mathbf{f}^s_{r,l}, \mathbf{f}^p_{r,l})$. To further verify our conjecture, we conduct strategy T8. It turns out that T8 has little difference with T7, validating that $\mathcal{L}_{ltrp}$ can effectively bridge the distance between $\mathbf{f}^s_{r,l}$ and $\mathbf{f}^p_{r,l}$.

4.4.3 How Does $\mathcal{L}_{ltrp-rn}$ Make Regions Semantically Discriminative? Sec. 4.4.1 and Fig. 2 qualitatively study the importance of $\mathcal{L}_{ltrp-rn}$ for MLRM. In this section, we quantify its effectiveness by comparing average intra-instance one-to-one region distances in each level: $\overline{D}_l = \mathbb{E}_{x\in\{s,p\}}\mathbb{E}_{r\neq r'} D(\mathbf{f}^x_{r,l}, \mathbf{f}^x_{r',l})$. The results are reported in Tab. 5. Without $\mathcal{L}_{ltrp-rn}$, the table shows that region features tend to converge to the same and $\overline{D}_l$ becomes small. This is mainly because the receptive field has covered the whole input image as the network depth grows[23]. As a result, all region features mainly contain global information. On the contrary, $\mathcal{L}_{ltrp-rn}$ can make different region features semantically discriminative and contain local information, therefore enlarging $\overline{D}_l$.

4.4.4 Case Study. We show top-5 retrievals and attended regions in Fig. 6. In general, MLRM is able to retrieve the images that share the same poses and categories with the query sketch. On QMUL-ChairV2, since all images and sketches are well aligned, the resultant attended regions are consistent in all cases. They are the top, the middle and the bottom of chairs (row (1)-(3)). On the contrary, attended regions on Sketchy differ from one another due to data misalignment and category variety. For example, the regions in row (5) are respectively the head, the neck and the body of a zebra (case 1), a rhino (case 3) and an elephant (case 4). These results demonstrate that MLRM is able to capture both geometrically and semantically discriminative regions and effectively perform region matching for images and sketches. We also conduct retrieval through the 128-dimensional positional embeddings. The resultant acc@1 is 0.8% on QMUL-ChairV2 and 1.0% on Sketchy. Considering that Sketchy is 40 times larger than QMUL-ChairV2, we can conclude that our positional embeddings based on geometry maps can well adapt to different poses and categories.

However, our method does not handle object size mismatch since the activated region size is fixed by τ. This is one reason for the top-1 failure of row (6), where the query sketch object is much larger than the paired image object. In our future work, we will investigate an approach that adaptively adjusts the region size. Another two reasons for the row (6) failure may be: **1)** The reflection in water in the paired image (case 3) is mistaken for a part of the crane; **2)** the first two cases share quite similar outlines with the query sketch.

5 CONCLUSION

To establish fine-grained correspondence between sketches and images, we propose a novel method named Multi-Level Region Matching (MLRM). MLRM consists of two modules: *Discriminative Region Extraction* (DRE) and *Region and Level Attention* (RLA). In DRE, we propose a LAMA structure to extract different attention maps to attend multi-level CNN feature maps. To ensure geometrical and semantic distinctiveness of the different attended regions, we explore a LAMA overlapping penalty and a local region-negative triplet loss. In RLA, we adopt a transformer-based attentive matching module to obtain attention weights for different regions and levels. To distinguish similar images, we propose a global region-negative triplet loss. Comprehensive experiments have demonstrated effectiveness of MLRM. In the future, we will seek to make region sizes self-adaptive to generalize MLRM to more complex cases. Also, we will further extend the idea to relevant tasks, *e.g.*, image matching and object localization.

Acknowledgement. The work is supported by Shanghai Municipal Science and Technology Major Project, China (2021SHZDZX0102), and Shanghai Municipal Science and Technology Key Project (Grant No. 20511100300), and National Science Foundation of China (61902247).

REFERENCES

[1] Hossein Azizpour and Ivan Laptev. 2012. Object Detection Using Strongly-Supervised Deformable Part Models. In *European Conference on Computer Vision (ECCV)*, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (Eds.). 836–849.

[2] Sreyasee Das Bhattacharjee, Junsong Yuan, Weixiang Hong, and Xiang Ruan. 2016. Query Adaptive Instance Search using Object Sketches. In *ACM Conference on Multimedia Conference, (ACM MM)*. 1306–1315.

[3] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. 2021. More Photos are All You Need: Semi-Supervised Learning for Fine-Grained Sketch Based Image Retrieval. In *Computer Vision and Pattern Recognition (CVPR)*.

[4] Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. 2022. Sketching without Worrying: Noise-Tolerant Sketch-Based Image Retrieval. In *Computer Vision and Pattern Recognition (CVPR)*.

[5] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. 2020. Sketch Less for More: On-the-Fly Fine-Grained Sketch-Based Image Retrieval. In *Computer Vision and Pattern Recognition (CVPR)*.

[6] Tu Bui, Leonardo Sampaio Ferraz Ribeiro, Moacir Ponti, and John P. Collomosse. 2018. Sketching Out The Details: Sketch-Based Image Retrieval Using Convolutional Neural Networks With Multi-Stage Regression. *Computers & Graphics (CAG)* 71 (2018), 77–87.

[7] Yangdong Chen, Zhaolong Zhang, Yanfei Wang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. 2022. AE-Net: Fine-Grained Sketch-Based Image Retrieval via Attention-Enhanced Network. *Pattern Recognition(PR)* (2022).

[8] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. 2010. An Evaluation of Descriptors for Large-Scale Image Retrieval from Sketched Feature Lines. *Computers & Graphics (CAG)* 34, 5 (2010), 482–498.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. 2010. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 32, 9 (2010), 1627–1645.

[10] Rui Hu and John Collomosse. 2013. A Performance Evaluation of Gradient Field HOG Descriptor for Sketch Based Image Retrieval. *Computer Vision and Image Understanding (CVIU)* (2013).

[11] Rui Hu, Tinghuai Wang, and John Collomosse. 2011. A Bag-of-Regions Approach to Sketch-Based Image Retrieval. In *IEEE International Conference on Image Processing (ICIP)*.

[12] Fei Huang, Yong Cheng, Cheng Jin, Yuejie Zhang, and Tao Zhang. 2017. Deep Multimodal Embedding Model for Fine-Grained Sketch-Based Image Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 929–932.

[13] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Computer Vision and Pattern Recognition (CVPR)*. 2261–2269.

[14] Toshikazu Kato, Takio Kurita, Nobuyuki Otsu, and Kyoji Hirata. 1992. A Sketch Retrieval Method for Full Color Image Database-Query by Visual Example. In *International Conference on Pattern Recognition (ICPR)*. 530–533.

[15] Jiangtong Li, Zhixin Ling, Li Niu, and Liqing Zhang. 2022. Zero-Shot Sketch-Based Image Retrieval with Structure-Aware Asymmetric Disentanglement. *Computer Vision and Image Understanding (CVIU)* 218 (2022), 103412.

[16] Ke Li, Kaiyue Pang, Yi-Zhe Song, Timothy Hospedales, Honggang Zhang, and Yichuan Hu. 2016. Fine-Grained Sketch-Based Image Retrieval: the Role of Part-Aware Attributes. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1–9.

[17] Ke Li, Kaiyue Pang, Yi Zhe Song, Timothy M. Hospedales, Tao Xiang, and Honggang Zhang. 2017. Synergistic Instance-Level Subspace Alignment for Fine-Grained Sketch-Based Image Retrieval. *IEEE Transactions on Image Processing* 26, 12 (2017), 5908–5921.

[18] Yi Li, Timothy M Hospedales, Yi-Zhe Song, and Shaogang Gong. 2014. Intra-category Sketch-Based Image Retrieval by Matching Deformable Part Models. In *British Machine Vision Conference (BMVC)*. 115.1–115.12.

[19] Hangyu Lin, Yanwei Fu, Peng Lu, Shaogang Gong, Xiangyang Xue, and Yu-Gang Jiang. 2019. TC-Net for iSBIR: Triplet Classification Network for Instance-Level Sketch Based Image Retrieval. In *ACM International Conference on Multimedia (ACM MM)*.

[20] Min Lin, Qiang Chen, and Shuicheng Yan. 2014. Network in Network. In *International Conference on Learning Representations (ICLR)*.

[21] Zhixin Ling, Zhen Xing, and Xiangdong Zhou. 2022. Conditional Stroke Recovery for Fine-Grained Sketch-Based Image Retrieval. In *European Conference on Computer Vision (ECCV)*.

[22] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. 2017. Deep Sketch Hashing: Fast Free-Hand Sketch-Based Image Retrieval. In *computer vision and pattern recognition (CVPR)*. 2862–2871.

[23] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard S. Zemel. 2016. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*. 4898–4906.

[24] Jinjie Mai, Meng Yang, and Wenfeng Luo. 2020. Erasing Integrated Learning: A Simple Yet Effective Approach for Weakly Supervised Object Localization. In *Computer Vision and Pattern Recognition (CVPR)*. 8766–8775.

[25] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. 2019. Generalising Fine-Grained Sketch-Based Image Retrieval. In *Computer Vision and Pattern Recognition (CVPR)*. 677–686.

[26] Kaiyue Pang, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. 2020. Solving Mixed-Modal Jigsaw Puzzle for Fine-Grained Sketch-Based Image Retrieval. In *Computer Vision and Pattern Recognition (CVPR)*.

[27] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. 2018. Deep Shape Matching. In *European Conference on Computer Vision (ECCV)*. 751–767.

[28] Jose M Saavedra, Juan Manuel Barrios, and S Orand. 2015. Sketch Based Image Retrieval Using Learned KeyShapes (LKS).. In *British Machine Vision Conference (BMVC)*. 164.1–164.11.

[29] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies. *ACM Transactions on Graphics (TOG)* (2016).

[30] Omar Seddati, Stéphane Dupont, and Mahmoudi Saïd. 2017. Quadruplet Networks for Sketch-Based Image Retrieval. In *ACM International Conference on Multimedia (ACM MM)*.

[31] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. 2018. Zero-Shot Sketch-Image Hashing. In *Computer Vision and Pattern Recognition (CVPR)*. 3598–3607.

[32] Jifei Song, Yi Zhe Song, Tao Xiang, Timothy Hospedales, and Ruan Xiang. 2016. Deep Multi-Task Attribute-Driven Ranking for Fine-Grained Sketch-Based Image Retrieval. In *British Machine Vision Conference (BMVC)*, Vol. 1. 3.

[33] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. 2017. Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval. In *International Conference on Computer Vision (ICCV)*. 5551–5560.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems (NIPS)*. 5998–6008.

[35] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming Ming Cheng, Yao Zhao, and Shuicheng Yan. 2017. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. In *Computer Vision and Pattern Recognition (CVPR)*. 1568–1576.

[36] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional Block Attention Module. In *European Conference on Computer Vision (ECCV)*, Vol. 11211. 3–19.

[37] Jiaqing Xu, Haifeng Sun, Qi Qi, Jingyu Wang, Ce Ge, Lejian Zhang, and Jianxin Liao. 2021. DLA-Net for FG-SBIR: Dynamic Local Aligned Network for Fine-Grained Sketch-Based Image Retrieval. In *ACM International Conference on Multimedia (ACM MM)*. 5609–5618.

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and tell: Neural Image Caption Generation with Visual Attention. In *International conference on machine learning (ICML)*. 2048–2057.

[39] Peng Xu, Qiyue Yin, Yonggang Qi, Yi-Zhe Song, Zhanyu Ma, Liang Wang, and Jun Guo. 2016. Instance-Level Coupled Subspace Learning for Fine-Grained Sketch-Based Image Retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 19–34.

[40] Wang Yanfei, Huang Fei, Zhang Yuejie, Feng Rui, Zhang Tao, and Fan Weiguo. 2019. Deep Cascaded Cross-Modal Correlation Learning for Fine-Grained Sketch-Based Image Retrieval. *Pattern Recognition(PR)* (2019).

[41] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. 2019. Towards Rich Feature Discovery with Class Activation Maps Augmentation for Person Re-Identification. In *Computer Vision and Pattern Recognition (CVPR)*. 1389–1398.

[42] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. 2016. Sketch Me That Shoe. In *Computer Vision and Pattern Recognition (CVPR)*.

[43] Xianlin Zhang, Mengling Shen, Xueming Li, and Fangxiang Feng. 2022. A Deformable CNN-Based Triplet Model for Fine-Grained Sketch-Based Image Retrieval. *Pattern Recognit. (PR)* 125 (2022), 108508.

[44] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. 2018. Adversarial Complementary Learning for Weakly Supervised Object Localization. In *Computer Vision and Pattern Recognition (CVPR)*. 1325–1334.

[45] Zhaolong Zhang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. 2020. Zero-Shot Sketch-Based Image Retrieval via Graph Convolution Network. In

AAAI Conference on Artificial Intelligence (AAAI), Vol. 34. 12943–12950.

[46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *Computer Vision and Pattern Recognition (CVPR)*. 2921–2929.

[47] Han Zhu, Mingsheng Long, Jianmin Wang, and Yue Cao. 2016. Deep Hashing Network for Efficient Similarity Retrieval. In *AAAI Conference on Artificial Intelligence (AAAI)*. 2415–2421.

[48] Ming Zhu, Chun Chen, Nian Wang, Jun Tang, and Wenxia Bao. 2019. Gradually Focused Fine-Grained Sketch-Based Image Retrieval. *PLoS ONE* (2019), e0217168.